ANALYZING AND MODELING ANALYTICAL VERB FORMS IN UZBEK BASED ON MORPHO-SYNTACTIC TAGS

Firuza Kamolovna Nurova,
Researcher, Department of Uzbek Linguistics, Bukhara State Pedagogical Institute Uzbekistan
Email: firuzaru870.@gmail.com

Abstract: *This study investigates the morphological and syntactic modeling of analytical verb forms in Uzbek through morpho-syntactic tagging techniques. Analytical constructions in Uzbek, especially auxiliary verb + participle formations, play a crucial role in expressing aspectual and modal nuances. The research builds a manually tagged corpus of 5,000 analytical verb instances using an extended tagset that distinguishes aspect, modality, tense, and polarity.*

Keywords: *Uzbek, NLP, analytical verb forms, morpho-syntactic tagging, corpus, auxiliary verbs.*

Morphological complexity in agglutinative languages such as Uzbek requires deep linguistic modeling for accurate computational processing. In particular, analytical verb forms – composed of a main verb and one or more auxiliary verbs – represent an important grammatical phenomenon that impacts sentence structure and meaning. While significant work has been done on simple verb morphology, little attention has been paid to analytical forms in machine-readable formats. The study aims to fill this gap by offering a novel approach to automatically recognize and tag these forms using modern machine learning methods. Such developments are critical for Uzbek to catch up with resource-rich languages in the field of NLP.

Uzbek language, being a member of the Turkic language family, exhibits a high degree of agglutinativity. Analytical verb forms in Uzbek often replace synthetic morphology to encode complex grammatical categories such as aspectual nuances, evidentiality, or modality. For instance, constructions like “ko‘rib chiqdi”, “borib keldi”, or “tushunib oldi” illustrate the multi-functionality of auxiliary verbs. These forms have also been influenced by contact with other Turkic and Persian languages, resulting in borrowing and structural convergence in analytic constructions. The challenge in identifying and modeling these forms computationally lies in distinguishing compositional analytical units from idiomatic or phraseological expressions. Therefore, a fine-grained morpho-syntactic analysis is critical for achieving high-performance natural language processing tools.

3. Methodology

Stage	Description
Data Collection	A sub-corpus of 5,000 sentences from the Uzbek National Corpus was selected containing analytical verbs.
Tagging Scheme	Tags were manually assigned for tense, aspect, modality, polarity, and verb combination patterns.
Models Used	CRF, BiLSTM+CRF, and BERT-base models were trained using annotated data.

The models were implemented using the HuggingFace Transformers library and trained on NVIDIA GPUs. Evaluation was performed using 10-fold cross-validation to ensure statistical significance.

Tagset Design Considerations

The tagset used in this research extends the traditional POS tagging framework by integrating functional and grammatical roles of auxiliaries. Each analytical form is annotated with tense/aspect (e.g., Progressive, Perfective, Habitual), modality (e.g., Necessitative, Possibility, Volition), evidentiality (e.g., Reported, Inferred), voice and polarity, and verb sequence pattern (e.g., MV+HV, Ger+HV+HV). This level of granularity enables more precise disambiguation and contributes to the enrichment of low-resource linguistic resources in Uzbek.

4. Results and Evaluation

Model	Accuracy	Precision	Recall
CRF	87.3%	85.1%	86.8%
BiLSTM+CRF	91.4%	90.2%	90.9%
BERT (fine-tuned)	**94.7%**	**94.1%**	**93.8%**

The tagging errors were mostly found in constructions involving multiple auxiliaries (e.g., “borib keldi edi”) or idiomatic expressions. Nevertheless, BERT-based models showed strong generalization across various analytical patterns.

Despite the success of BERT-based models, certain types of constructions remained difficult to classify correctly: - Discontinuous verb combinations such as “...kelib, yana borib keldi”.- Elliptical constructions, where auxiliary components are

omitted but implied. - Idiomatic expressions (e.g., “ko‘z yumdi” meaning “passed away”) that require contextual semantic understanding beyond syntactic patterns. Future modeling efforts could benefit from integrating semantic role labeling and dependency parsing to enhance analytical verb recognition.

5. Application and Implications. The resulting tagger can be integrated into larger Uzbek NLP pipelines for:

- Morphological Analysis
- Machine Translation
- Syntactic Parsing
- Voice Assistant Systems for Uzbek

Furthermore, this tagset can serve as a blueprint for other Turkic languages with similar verb formation patterns, supporting cross-lingual NLP.

Corpus Expansion and Semi-Automated Annotation Tools

To scale the project, a semi-automated annotation pipeline is under development. This system relies on a hybrid rule-based and ML-based architecture to pre-tag potential analytical constructions, which are then verified by human annotators. Additionally, plans include expanding the corpus with dialectal and domain-specific texts (journalistic, literary, medical), thereby increasing diversity and robustness of the tagger.

This research demonstrates that analytical verb forms in Uzbek can be effectively modeled using modern tagging and deep learning techniques. The annotated dataset and models will aid in developing robust Uzbek NLP pipelines and can be adapted to other Turkic languages with similar analytical structures. Future work includes expanding the corpus size and integrating syntactic treebanks for deeper analysis.

REFERENCES:

1. Abdurakhmonov, N. (2017). Modeling analytic forms of verb in Uzbek as stage of morphological analysis in machine translation. Journal of Social Sciences and Humanities Research, 5(03), 89-100.

2. Abdurakhmonova N. O’zbek tili korpusini morfologik teglashda FST texnologiyasi tatbiqi. International Journal of Art & Design Education 2021; 4: 319–326.

3. Maksud Sharipov, Ulugbek Salaev, Gayrat Matlatipov. IMPLEMENTED STEMMING ALGORITHMS BASED ON FINITE STATE MACHINE FOR UZBEK VERBS |COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS. 2022 http://compling.navoiy-uni.uz/index.php/conferences/article/view/6 (May 30, 2022, date last accessed)

4. Maqsud Sharipov. Uzbek_POS_tag_list/Uzbek POS tag list.pdf at mainMaksudSharipov/Uzbek_POS_tag_list·GitHub.2020

https://github.com/MaksudSharipov/Uzbek_POS_tag_list/blob/main/Uzbek%20POS%20tag%20list.pdf (May 28, 2022, date last accessed).

5. Марчук Ю. Компьютерная лингвистика. - Москва: АСТ Восток-Запад, 2007. – 174 с

6. Mengliev, B., Shahabitdinova, S., Khamroeva, S., Gulyamova, S., & Botirova, A. (2020). The Morphological Analysis and Synthesis of Word Forms in the Linguistic Analyzer.